PAC-Bayesian Bounds for Sparse Regression Estimation with Exponential Weights

Pierre Alquier, Karim Lounici

PAC-BAYESIAN BOUNDS FOR SPARSE REGRESSION ESTIMATION WITH EXPONENTIAL WEIGHTS

PIERRE ALQUIER, KARIM LOUNICI

Abstract. We consider the sparse regression model where the number of parameters p is larger than the sample size n. The difficulty when considering high-dimensional problems is to propose estimators achieving a good compromise between statistical and computational performances. The BIC estimator for instance performs well from the statistical point of view [11] but can only be computed for values of p of at most a few tens. The Lasso estimator is solution of a convex minimization problem, hence computable for large value of p. However stringent conditions on the design are required to establish fast rates of convergence for this estimator. Dalalyan and Tsybakov [19] propose a method achieving a good compromise between the statistical and computational aspects of the problem. Their estimator can be computed for reasonably large p and satisfies nice statistical properties under weak assumptions on the design. However, [19] proposes sparsity oracle inequalities in expectation for the empirical excess risk only. In this paper, we propose an aggregation procedure similar to that of [19] but with improved statistical performances. Our main theoretical result is a sparsity oracle inequality in probability for the true excess risk for a version of exponential weight estimator. We also propose a MCMC method to compute our estimator for reasonably large values of p.

MSC 2000 subject classification: Primary: 62J07; Secondary: 62J05, 62G08, 62F15, 62B10, 68T05.

Key words and phrases: Sparsity Oracle Inequality, High-dimensional Regression, Exponential Weights, PAC-Bayesian, RJMCMC

1. Introduction

We observe n independent pairs $(X_1, Y_1), ..., (X_n, Y_n) \in \mathcal{X} \times \mathbb{R}$ (where $\mathcal{X}$ is any measurable set) such that

$$Y_i = f(X_i) + W_i,\ 1 \leqslant i \leqslant n, \tag{1.1}$$

where $f : \mathcal{X} \to \mathbb{R}$ is the unknown regression function and the noise variables $W_1, \ldots, W_n$ are independent of the design $(X_1, \ldots, X_n)$ and such that $\mathbb{E}W_i = 0$ and $\mathbb{E}W_i^2 \leqslant \sigma^2$ for some known $\sigma^2 > 0$ and any $1 \leqslant i \leqslant n$. The distribution of the sample is denoted by $\mathbb{P}$, the corresponding expectation is denoted by $\mathbb{E}$. For any function $g : \mathcal{X} \to \mathbb{R}$ define $\|g\|_n = \left(\sum_{i=1}^n g(X_i)^2/n\right)^{1/2}$ and $\|g\| = \left(\mathbb{E}\|g\|_n^2\right)^{1/2}$. Let $\mathcal{F} = \{\phi_1, \ldots, \phi_p\}$ be a set—called dictionary—of functions $\phi_j : \mathcal{X} \to \mathbb{R}$ such that $\|\phi_j\| = 1$ for any j (this assumption can be relaxed). For any $\theta \in \mathbb{R}^p$ define $f_\theta = \sum_{j=1}^p \theta_j \phi_j$ and the risk

$$R(\theta) = E\left[\frac{1}{n}\sum_{i=1}^n \left(Y_i' - f_\theta(X_i')\right)^2\right],$$

where $\{(X_1', Y_1'), \ldots, (X_n', Y_n')\}$ is an independent replication of the sample. Let us choose $\overline{\theta} \in \arg\min_{\theta \in \mathbb{R}^p} R(\theta)$. Note that the minimum may not be unique but since

Date: September 14, 2010.

we consider in this paper the prediction problem, we do not need to deal with the identifiability question.

It is a known fact that the least-square estimator $\hat{\theta}_n^{LSE} \in \arg\min_{\theta\in\Theta} r(\theta)$ performs poorly in high-dimension $p > n$. Indeed, consider for instance the deterministic design case with i.i.d. noise variables $\mathcal{N}(0, \sigma^2)$ and a full-rank design matrix, then $\hat{\theta}^{LSE}$ satisfies

$$\mathbb{E}\left[\|f_{\hat{\theta}_n^{LSE}} - f\|_n^2\right] - \|f_{\overline{\theta}} - f\|_n^2 = \sigma^2.$$

In the same context, assume now there exists a vector $\overline{\theta} \in \arg\min_{\theta\mathbb{R}^p} R(\theta)$ with a number of nonzero coordinates $p_0 \leq n$. If the indices of these coordinates are known, then we can construct an estimator $\hat{\theta}_n^0$ such that

$$\mathbb{E}\left[\|f_{\hat{\theta}_n^0} - f\|_n^2\right] - \|f_{\overline{\theta}} - f\|_n^2 = \sigma^2 \frac{p_0}{n}.$$

The estimator $\hat{\theta}_n^0$ is called oracle estimator since the set of indices of the nonzero coordinates of $\overline{\theta}$ is unknown in practice. The issue is now to build an estimator, when the set of nonzero coordinates of $\overline{\theta}$ is unknown, with statistical performances close to that of the oracle estimator $\hat{\theta}_n^0$.

A possible approach is to consider solutions of penalized empirical risk minimization problems:

$$\hat{\theta}_{pen} \in \arg\min_{\theta\in\Theta} \left\{ \frac{1}{n}\sum_{i=1}^{n}\Big(Y_i - f_\theta(X_i)\Big)^2 + pen(\theta) \right\},$$

where the penalization $pen(\theta)$ is proportional to the number of nonzero components of θ as for instance C_p, AIC and BIC criteria [39, 1, 47]. Bunea, Tsybakov and Wegkamp [11] established for the BIC estimator $\hat{\theta}_n^{BIC}$ the following non-asymptotic sparsity oracle inequality. For any $\epsilon > 0$ there exists a constant $C(\epsilon) > 0$ such that for any $p \geqslant 2, n \geqslant 1$ we have

$$\mathbb{E}\left[\|f_{\hat{\theta}_n^{BIC}} - f\|_n^2\right] \leqslant (1+\epsilon)\|f_{\overline{\theta}} - f\|_n^2 + C(\epsilon)\sigma^2 \frac{p_0}{n} \log\left(\frac{ep}{p_0 \vee 1}\right).$$

Despite good statistical properties, these estimators can only be computed in practice for p of the order at most a few tens since they are solutions of non-convex optimization problems.

Considering convex penalty function leads to computationally feasible optimization problems. A popular example of convex optimization problem is the Lasso estimator (cf. Frank and Friedman [25], Tibshirani [50], and the parallel work of Chen *et al* on basis pursuit [17]) with the penalty term $pen(\theta) = \lambda|\theta|_1$, where $\lambda > 0$ is some regularization parameter and, for any integer $d \geq 2$, real $q > 0$ and vector $z \in \mathbb{R}^d$ we define $|z|_q = (\sum_{j=1}^d |\theta_j^q|)^{1/q}$ and $|z|_\infty = \max_{1\leq j\leq d} |\theta_j|$. Several algorithms allow to compute the LASSO for very large p, one of the most popular is known as LARS, introduced by Efron *et al* [24]. However, the Lasso estimator requires strong assumptions on the matrix $A = (\phi_j(X_i))_{1\leqslant i\leqslant n, 1\leqslant j\leqslant p}$ to establish fast rates of convergence results. Bunea, Tsybakov and Wegkamp [10] assume a mutual coherence condition on the dictionary. Bickel, Ritov and Tsybakov [8] and Koltchinskii [34] established sparsity oracle inequalities for the Lasso under a restricted eigenvalue condition. Candès and Tao [13] proposed the Dantzig Selector which is related to the Lasso estimator and suffers from the same restrictions. See for example Bickel, Ritov and Tsybakov [8] for more details. Several alternative penalties were recently considered. Zou [56] proposed the adaptive LASSO which is the solution of a penalized empirical risk minimization problem with the penalty $pen(\theta) = \lambda \sum_{j=1}^p \frac{1}{|\hat{w}_j|}|\theta_j|$ where $\hat{w}$ is an a priori estimator. Zou and Hastie [57]

proposed the elastic net with the penalty $pen(\theta) = \lambda_1|\theta|_1 + \lambda_2|\theta|_2^2$, $\lambda_1, \lambda_2 > 0$. Meinshausen and Bühlmann [43] and Bach [6] considered bootstrapped LASSO. See also Ghosh [28] or Cai, Xu and Zhang [12] for more alternatives to the LASSO. All these methods were motivated by their superior performances over the LASSO either from the theoretical or the practical point of view. However, strong assumptions on the design are still required to establish the statistical properties of these methods (when such results exist). A recent paper by van de Geer and Bühlmann [52] provides a complete survey and comparison of all these assumptions.

Simultaneously, the PAC-Bayesian approach for regression estimation was developed by Audibert [4, 5] and Alquier [2, 3], based on previous works in the classification context by Catoni [14, 15, 16], Mc Allester [42], Shawe-Taylor and Williamson [49], see also Zhang [55] in the context of density estimation. This framework is very well adapted for studying the excess risk $R(\cdot) - R(\overline{\theta})$ in the regression context since it requires very weak conditions on the dictionary. However, the methods of these papers are not computationally feasible when p becomes large. Dalalyan and Tsybakov [19, 20, 21, 22] propose an exponential weights procedure related to the PAC-Bayesian approach with good statistical and computational performances. However they consider deterministic design, establishing their statistical result only for the empirical excess risk instead of the true excess risk $R(\cdot) - R(\overline{\theta})$.

In this paper, we propose to study two exponential weights estimation procedures. The first one is an exponential weights combination of the least squares estimators in all the possible sub-models. This estimator was initially proposed by Leung and Barron [36] in the deterministic design setting. Note that in the literature on aggregation, the elements of the dictionary are often preliminary arbitrary estimators computed from a frozen fraction of the initial sample so that these estimators are considered as deterministic functions, the aggregate is then computed using this dictionary and the remaining data. This scheme is referred to as 'data splitting'. See for instance Dalalyan and Tsybakov [20, 21] and Yang [54]. Leung and Barron [36] proved that data splitting is not necessary in order to aggregate least squares estimators and raised the question of computation of this estimator in high dimension. In this paper we explicit the oracle inequality satisfied by this estimator in the high-dimensional case and tackle the computational question. For the second procedure, the design is assumed to be random. We use the PAC-Bayesian techniques of Catoni [16] to build an estimator satisfying a sparsity oracle inequality for the true excess risk. Then we propose computationally efficient Monte Carlo algorithms to compute both estimators. Our algorithms are inspired from the computational Bayesian theory, see the monograph of Marin and Robert [40] for an introduction to Monte Carlo algorithms in Bayesian theory. More specifically, the Bayesian point of view for the variable selection problem was considered in several papers: George [26], George and McCulloch [27], West [53], Jiang [32], Cui and George [18], Bogdan *et al* [9], Liang *et al* [37], Scott and Berger [48] among others. See in particular [27, 44] for the algorithmic aspects of Monte Carlo techniques. In this paper, we use Hastings Metropolis algorithm to compute our first estimator and we implement a version of the RJMCMC ("Reversible Jump Markov Chain Monte Carlo") method proposed by Green [29] to compute our second estimator. Note that in a work parallel to ours, Rigollet and Tsybakov [46] consider exponentially weighted aggregates with discrete priors and suggest another version of the Metropolis-Hastings algorithm to compute their estimator.

The paper is organized as follows. In Section 2 we define a general aggregation procedure and derive a sparsity oracle inequality in the deterministic design case. In Section 3 the design can be either deterministic or random. We propose a modification of the first aggregation procedure for which we can establish a sparsity

oracle inequality in probability for the true excess risk. Section 4 is devoted to the RJMCMC algorithm used to effectively implement our estimators. In Section 5 we carry out a simulation study and compare the performances of our methods with the Lasso. Finally Section 6 contains all the proofs of our results.

2. Sparsity Oracle Inequality in Expectation

Throughout this section, we assume that the design is deterministic and the noise variables $W_1, \dots, W_n$ are i.i.d. gaussian $N(0, \sigma^2)$.

For any $J \subset \{1, ..., p\}$ and $K > 0$ define

$$\Theta(J) = \left\{ \theta \in \mathbb{R}^p : \quad \forall j \notin J, \quad \theta_j = 0 \right\}, \tag{2.1}$$

and

$$\Theta_K(J) = \left\{ \theta \in \mathbb{R}^p : \quad |\theta|_1 \leq K \quad \text{and} \quad \forall j \notin J, \quad \theta_j = 0 \right\}. \tag{2.2}$$

For the sake of simplicity we will write $\Theta_K = \Theta_K(\{1, ..., p\})$.

For any subset $J \subset \{1, ..., p\}$ define

$$\hat{\theta}_J \in \arg\min_{\theta \in \Theta(J)} r(\theta),$$

where $r(\theta) = \frac{1}{n} \sum_{i=1}^n \Big(Y_i - f_\theta(X_i) \Big)^2 = \|Y - f_\theta\|_n^2$ with $Y = (Y_1, ..., Y_n)^T$. Denote by $\mathcal{P}_n(\{1, \dots, p\})$ the set of all subsets of $\{1, \dots, p\}$ containing at most n elements. The aggregate $\hat{f}_n$ is defined as follows

$$\hat{f}_n = f_{\hat{\theta}_n}, \quad \hat{\theta}_n = \hat{\theta}_n(\lambda, \pi) \stackrel{\triangle}{=} \frac{\sum_{J \in \mathcal{P}_n(\{1,...,p\})} \pi_J e^{-\lambda \left(r(\hat{\theta}_J) + \frac{2\sigma^2 |J|}{n} \right)} \hat{\theta}_J}{\sum_{J \in \mathcal{P}_n(\{1,...,p\})} \pi_J e^{-\lambda \left(r(\hat{\theta}_J) + \frac{2\sigma^2 |J|}{n} \right)}} \tag{2.3}$$

where $\lambda > 0$ is the temperature parameter, π is the prior probability distribution on $\mathcal{P}(\{1, \dots, p\})$, the set of all subsets of $\{1, ..., p\}$, that is, for any $J \in \{1, ..., p\}$, $\pi_J \geq 0$ and $\sum_{J \in \mathcal{P}(\{1,...,p\})} \pi_J = 1$. In Section 4 we show that the estimator $\hat{\theta}_n$ can be computed in reasonable time even when p is large using a MCMC scheme, namely, Hastings-Metropolis algorithm. The parameters π and λ must be tuned in a suitable way. The choice of π is discussed below. The choice of the temperature parameter λ is discussed in Section 5.

We now state the main results of this section.

Proposition 1. *Assume that the noise variables $W_1, \dots, W_n$ are i.i.d. $N(0, \sigma^2)$. Then the aggregate $\hat{\theta}_n$ defined by (2.3) with $0 < \lambda \leqslant \frac{n}{4\sigma^2}$ satisfies*

$$\mathbb{E}\left[r(\hat{\theta}_n) \right] \leqslant \min_{J \in \mathcal{P}_n(\{1,...,p\})} \left\{ \mathbb{E}[r(\hat{\theta}_J)] + \frac{1}{\lambda} \log\left(\frac{1}{\pi_J} \right) \right\}. \tag{2.4}$$

Proposition 1 holds true for any prior π and is due to Leung and Barron [36]. In what follows, we exploit this result in order to establish a sharp sparsity oracle inequality for the aggregation procedure (2.3). We suggest the following prior. Fix $\alpha \in (0, 1)$. Define π as follows

$$\pi_J = \frac{\alpha^{|J|}}{\sum_{j=0}^n \alpha^j} \binom{p}{|J|}^{-1}, \quad \forall J \in \mathcal{P}(\{1, ..., p\}). \tag{2.5}$$

We have the following theorem

Theorem 1. *Assume that the noise variables $W_1, \dots, W_n$ are i.i.d. $N(0, \sigma^2)$. Then the aggregate $\hat{f}_n = f_{\hat{\theta}_n}$, with $\lambda = \frac{n}{4\sigma^2}$ and π taken as in (2.5), satisfies*

$$\mathbb{E}\left[\|\hat{f}_n - f\|_n^2\right] \leqslant \min_{\theta \in \mathbb{R}^p} \left\{ \|f_\theta - f\|_n^2 + \frac{\sigma^2 |J(\theta)|}{n} \left(4 \log\left(\frac{pe}{|J(\theta)|\alpha} \right) + 1 \right) + \frac{4\sigma^2 \log\left(\frac{1}{1-\alpha}\right)}{n} \right\}, \tag{2.6}$$

where for any $\theta \in \mathbb{R}^p$ $J(\theta) = \{j : \theta_j \neq 0\}$.

Tsybakov [51] introduced the notion of optimal rate of aggregation adapting existing tools from the minimax theory. In particular, the rate derived in Theorem 1 is the optimal rate of sparse linear aggregation and does not depend on the magnitude of the nonzero components of θ. This result can be compared with the sparsity oracle inequalities established in [8, 13, 19] where the rates becomes very large if the nonzero components of θ take large values.

3. Sparsity Oracle Inequality in Probability

From now on, the design can be either deterministic or random. We make the following mild assumption:

$$L = \max_{1 \leq j \leq M} \|\phi_j\|_\infty < \infty.$$

We assume in this section that the noise variables are subgaussian. More precisely we have the following condition.

Assumption 1. The noise variables $W_1, \dots, W_n$ are independent and independent of $X_1, \dots, X_n$. We assume also that there exist two known constants $\sigma > 0$ and $\xi > 0$ such that

$$\mathbb{E}(W_i^2) \leq \sigma^2$$
$$\forall k \geq 3, \quad \mathbb{E}(|W_i|^k) \leq \sigma^2 k! \xi^{k-2}.$$

The estimation method is a version of the Gibbs estimator introduced by Catoni [15, 16]. Fix $K \geq 1$ and $c > 0$. First we define the prior probability distribution as follows. For any $J \subset \{1, ..., p\}$ let u_J denote the uniform measure on $\Theta_{K+c}(J)$. We define

$$\mathrm{m}(d\theta) = \sum_{J \subset \{1,\dots,p\}} \pi_J \mathrm{u}_J(d\theta)$$

with π taken as in (2.5).

We are now ready to define our estimator. For any $\lambda > 0$ we consider the probability measure $\tilde{\rho}_\lambda$ admitting the following density w.r.t. the probability measure m

$$\frac{d\tilde{\rho}_\lambda}{d\mathrm{m}}(\theta) = \frac{e^{-\lambda r(\theta)}}{\int_{\Theta_K} e^{-\lambda r} d\mathrm{m}}. \tag{3.1}$$

The aggregate $\tilde{f}_n$ is defined as follows

$$\tilde{f}_n = f_{\tilde{\theta}_n}, \quad \tilde{\theta}_n = \tilde{\theta}_n(\lambda, \mathrm{m}) = \int_{\Theta_K} \theta \tilde{\rho}_\lambda(d\theta). \tag{3.2}$$

The practical computation of $\tilde{\theta}_n$ is discussed in Section 4.

Define

$$\mathcal{C}_1 = \left[8\sigma^2 + (2\|f\|_\infty + L(2K + c))^2\right] \vee \left[8[\xi + (2\|f\|_\infty + L(2K + c))]L(2K + c)\right].$$

We can now state the main result of this section.

Theorem 2. *Let Assumption 1 be satisfied. Take $K > 1$, $c = n^{-1}$ and $\lambda = \lambda^* = \frac{n}{2\mathcal{C}_1}$. Assume that $\arg\min_{\theta\in\mathbb{R}^p} R(\theta) \cap \Theta_K \neq \emptyset$. Then we have, for any $\varepsilon \in (0,1)$ and any $\bar{\theta} \in \arg\min_{\theta\in\mathbb{R}^p} R(\theta) \cap \Theta_K$, with probability at least $1-\epsilon$,*

$$R(\tilde{\theta}_n) \leq R(\bar{\theta}) + \frac{3L^2}{n^2} + \frac{8\mathcal{C}_1}{n}\Bigg[|J(\bar{\theta})|\log(K+c) + \left(|J(\bar{\theta})|\log\left(\frac{enp}{\alpha|J(\bar{\theta})|}\right) + \log\left(\frac{2}{\varepsilon(1-\alpha)}\right)\right)\Bigg].$$

This theorem improves upon previous results on the following points:

- Our result holds under mild conditions on the dictionary while majority of the results under the sparsity scenario impose stringent conditions on the dictionary (see, e.g, [6, 8, 13]).
- we state a sparsity oracle inequality in probability for the risk $R(\cdot)$ whereas previous results concern either the empirical risk or are given in expectation [19, 33, 38].
- Unlike mirror averaging or progressive mixture rules, satisfying similar inequalities in expectation, our estimator does not involve an averaging step. As a consequence, its computational complexity is significantly reduced as compared to those procedures with averaging step.

The choice $\lambda = \lambda^*$ comes from the optimization of a (rather pessimistic) upper bound on the risk R (see Inequality (6.7) in the proof of this theorem, page 18). However this choice is not necessarily the best choice in practice even though it gives the good order of magnitude for λ. Section 5 illustrates this point. The practitioner may use cross-validation to properly tune the temperature parameter.

4. Practical computation of the estimator

Practical computation of $\tilde{\theta}_n$ and $\hat{\theta}_n$, for a given temperature $\lambda > 0$, is delicate. Indeed, exact computation of these estimators requires considering all subsets $\Theta_K(J)$, for any $J \subset \{1,...,p\}$. Since there are 2^p such subsets, exact computation of $\hat{\theta}_n$ or $\tilde{\theta}_n$ is not feasible for large p. However, since our estimators are defined as expectations of posterior distributions, we can approximate them via Monte Carlo computation. There exists an extensive literature on Monte Carlo computational methods, especially in Bayesian statistics where estimators can sometimes be expressed as expectations of posterior distribution, see e.g. Marin and Robert [40]. A standard Markov Chains Monte Carlo (MCMC) algorithm such as Hastings-Metropolis can be used to compute $\hat{\theta}_n$ since the prior π used to define this estimator is a discrete probability distribution. The computation of $\tilde{\theta}_n$ is more delicate. Indeed $\tilde{\rho}_\lambda$ is absolutely continuous w.r.t. the measure $\mathrm{m}(d\theta)$ which involves a mixture of Lebesgue measures on spaces of different dimensions. A way to proceed with such measures was proposed by Green [29] under the name "Reversible Jump Markov Chain Monte Carlo", RJMCMC, and applied successfully in various problems of model selection like multiple change-point problems, image segmentation and partition models in [29] or selection of the number of components in a mixture model in Green and Richardson [30]. We propose to adapt this procedure to our setting to compute $\tilde{\theta}_n$.

4.1. Computation of $\hat{\theta}_n$ via Hastings-Metropolis sampling.

In this subsection, we write a particular form of Hastings-Metropolis algorithm that will allow to

compute any estimator of the form

$$\hat{\theta}^{(w)} = \sum_{J \in \mathcal{P}_n(\{1,\dots,p\})} w_J \hat{\theta}_J$$

where we already defined

$$\hat{\theta}_J \in \arg\min_{\theta \in \Theta(J)} r(\theta),$$

and

$$w_J \geq 0, \ \ \forall J \subset \{1,...,p\} \quad \text{and} \quad \sum_{J \in \mathcal{P}(\{1,\dots,p\})} w_J = 1.$$

Hastings-Metropolis algorithm starts from an arbitrary value (say $J^{(0)} = \emptyset$), a simple transition kernel $k(\cdot,\cdot)$ on the set $\{J \subset \{1,...,p\}, |J| \leq n\}$ and updates $J^{(t)}$ to $J^{(t+1)}$ using the following scheme:

- draw $I^{(t)}$ from $k(J^{(t)},\cdot)$;
- take

$$J^{(t+1)} = \begin{cases} I^{(t)} & \text{with probability } \alpha(J^{(t)}, I^{(t)}) = \min\left(1, \frac{w_{I^{(t)}} k(I^{(t)}, J^{(t)})}{w_{J^{(t)}} k(J^{(t)}, I^{(t)})}\right) \\ J^{(t)} & \text{with probability } 1 - \alpha(J^{(t)}, I^{(t)}). \end{cases}$$

We stop after T steps and compute our estimator as the mean of the $\hat{\theta}_{J^{(t)}}$:

$$\hat{\theta}^{w,T,bo} = \frac{1}{T - bo + 1} \sum_{t=bo}^{T} \hat{\theta}_{J^{(t)}}$$

where, as usual in MCMC methods, we remove the *bo* first simulations (*burn-in* period).

This algorithm ensures that $(J^{(t)})_t$ is a Markov chain with invariant probability distribution $(w_J)_J$ (see Marin and Robert [40]). Here, the set $\{J \subset \{1,...,p\}, |J| \leq n\}$ being finite, we just have to remark that the chain is irreducible and aperiodic to obtain the convergence of $\hat{\theta}^{w,T,bo}$ to $\hat{\theta}^{(w)}$ (in probability).

In practice, we use the following kernel k:

$$k(J,\cdot) = k_+(J,\cdot)\mathbb{1}_{\{|J|=0\}} + \frac{k_+(J,\cdot) + k_-(J,\cdot)}{2}\mathbb{1}_{\{0<|J|<n\}} + k_-(J,\cdot)\mathbb{1}_{\{|J|=n\}}$$

where $k_+(\cdot,\cdot)$ and $k_-(\cdot,\cdot)$ are two kernels that we define now. The kernel k_+ adds an element to J whereas k_- removes one element from J. When we try to add an element, it is reasonable to consider first features that are the most correlated with the current residual. Similarly when we try to remove one element, we give priority the feature with the smallest coefficients in absolute value.

Formally, we choose some parameter $\zeta > 0$. We put, for $j \notin J$:

$$k_+(J, J \cup \{j\}) = \frac{e^{\zeta |c_j|}}{\sum_{h \notin J} e^{\zeta |c_h|}}$$

where c_j is the coefficient of linear correlation between $\left(Y_i - f_{\hat{\theta}_J}(X_i)\right)_{1 \leq i \leq n}$ and $(\phi_j(X_i))_{1 \leq i \leq n}$. And, for $j \in J$:

$$k_-(J, J \setminus \{j\}) = \frac{e^{-\zeta \left|(\hat{\theta}_J)_j\right|}}{\sum_{h \in J} e^{-\zeta \left|(\hat{\theta}_J)_h\right|}}.$$

Remark 1. The rationale behind our choice for $k_+(\cdot,\cdot)$ is the following. If $\zeta = 0$, then the above algorithm adds a new coordinate to the model uniformly at random among the set of unused coordinates. This procedure, although reasonable in theory, is not efficient in practice when p becomes large. Indeed for large p (say $p = 10^4$ and $p_0 = 2$) a large number of steps can be necessary before we select an

interesting coordinate. On the other hand, for large ζ the above procedure selects at each step the coordinate to incorporate to the model in a greedy way, that is the coordinate corresponding to the most correlated element of the dictionary with the current residual. For more details on *greedy* algorithms, see Barron *et al* [7, 31]. This procedure is computationally efficient and performs better in high-dimension than the former one with $\zeta = 0$. However it can sometimes be trapped in a local minimum. Therefore using an intermediate value for ζ seems to give a good balance in practice between these two extreme cases.

4.2. **RJMCMC algorithm and computation of $\tilde{\theta}_n$.** The RJMCMC algorithm proposed by Green [29] is an application of the Hastings-Metropolis to the case of a measure absolutely continuous with relation to a more sophisticated distribution (in our case m). We use here this method to compute $\tilde{\theta}_n = \int_{\Theta_K} \theta \tilde{\rho}_\lambda(d\theta)$. We start from $\theta^{(0)} = 0$ and then, at each step, update $\theta^{(t)}$ to $\theta^{(t+1)}$ using the transition kernel. Note that we need to define a kernel k admitting a density w.r.t. the measure m. For the sake of simplicity, we denote by $\tilde{\rho}_\lambda(\cdot)$ the measure $\tilde{\rho}_\lambda$ as well as its density with respect to m, this is a standard convention in the MCMC literature. We define now the RJMCMC algorithm:

- draw $\tau^{(t)}$ from $k(\theta^{(t)}, \cdot)$;
- take

$$\vartheta^{(t)} = \begin{cases} \tau^{(t)} & \text{with proba. } \alpha(\theta^{(t)}, \tau^{(t)}) = \min\left(1, \frac{\tilde{\rho}_\lambda(\tau^{(t)}) k(\tau^{(t)}, \theta^{(t)})}{\tilde{\rho}_\lambda(\tau^{(t)}) k(\theta^{(t)}, \tau^{(t)})}\right) \\ \theta^{(t)} & \text{with proba. } 1 - \alpha(\theta^{(t)}, \tau^{(t)}) \end{cases} ;$$

- draw $\theta^{(t+1)}$ from the distribution $\tilde{\rho}_\lambda(d\theta | \theta \in \Theta(J(\vartheta^{(t+1)})))$.

This algorithm ensures that $(\theta^{(t)})_t$ is a Markov chain with invariant probability distribution ρ_λ, see [40]. The last step ensures that we can make a move inside the current model even if the model change was rejected by the Hastings-Metropolis step.

We define now the kernel k:

$$k(\theta, \cdot) = k_+(\theta, \cdot) \mathbb{1}_{\{J(\theta) = \emptyset\}} + \frac{k_+(\theta, \cdot) + k_-(\theta, \cdot)}{2} \mathbb{1}_{\{0 < |J(\theta)| < n\}} + k_-(\theta, \cdot) \mathbb{1}_{\{|J(\theta)| = n\}}$$

where, for some $\zeta > 0$,

$$k_+(\theta, d\theta') = \sum_{j \notin J(\theta)} \frac{e^{\zeta |c_j(\theta)|}}{\sum_{h \notin J(\theta)} e^{\zeta |c_h(\theta)|}} \tilde{\rho}_\lambda(d\theta' | \theta' \in \Theta(J(\theta) \cup \{j\}))$$

and

$$k_-(\theta, d\theta') = \sum_{j \in J(\theta)} \frac{e^{-\zeta |\theta_j|}}{\sum_{h \in J(\theta)} e^{-\zeta |\theta_h|}} \tilde{\rho}_\lambda(d\theta' | \theta' \in \Theta(J(\theta) \setminus \{j\})),$$

where $c_j(\theta)$ is the coefficient of linear correlation between $(Y_i - f_\theta(X_i))_{1 \le i \le n}$ and $(\phi_j(X_i))_{1 \le i \le n}$.

5. Simulations

We compare in this section the exponential weights estimators $\hat{\theta}_n$ and $\tilde{\theta}_n$ to the LASSO [50] and the Elastic Net [57] on a toy example.

5.1. **Description of the experiments.** We consider variants of the toy example in [50]:

$$\forall i \in \{1, \dots, n = 20\}, \quad Y_i = \langle \theta, X_i \rangle + \varepsilon_i$$

with $X_i \in \mathcal{X} = \mathbb{R}^p$, $\bar{\theta} \in \mathbb{R}^p$ and the ε_i are i.i.d. from a gaussian distribution with mean 0 and standard deviation σ.

The X_i's are i.i.d. (and independent from the ε_i) and drawn from the gaussian distribution with zero mean and variance matrix:

$$\Sigma(\rho) = \left(\rho^{|i-j|}\right)_{\substack{i \in \{1,\dots,p\} \\ j \in \{1,\dots,p\}}}$$

for some $\rho \in [0, 1)$. Note that Tibshirani's toy example is set with $p = 8$ whereas we will consider here $p \in \{8, 30, 100, 1000\}$.

Define the regression vector as either

$$\bar{\theta} = (3, 1.5, 0, 0, 2, 0, 0, 0, 0, 0, ...)$$
$$\text{or } \bar{\theta} = (5e^{-1}, 5e^{-2}, 5e^{-3}, 5e^{-4}, ...)$$

corresponding respectively to a «sparse situation», $|J(\bar{\theta})| = 3$ whatever is the value of p, and to an approximately sparse situation, with $|J(\bar{\theta})| = p$ but $\bar{\theta}$ well approximated in the l_1 norm by a vector $\theta' \in \mathbb{R}^p$ with $|J(\theta')| \ll p$.

We will take σ^2 respectively equal to 1 («low noise situation») and 3 («noisy case»); the value of ρ is fixed to 0.5. We fix $\zeta = 2$ in the RJMCMC algorithm, $\alpha = 1/10$, $T = 12000$ and $bo = 2000$.

The LASSO and the Elastic Net are defined respectively by

$$\hat{\theta}_n^L = \hat{\theta}_n^L(\mu) = \arg\min_{\theta \in \mathbb{R}^p} \left[r(\theta) + \mu |\theta|_1 \right], \ \mu > 0$$

and

$$\hat{\theta}_n^{EN} = \hat{\theta}_n^{EN}(\mu, \gamma) = \arg\min_{\theta \in \mathbb{R}^p} \left[r(\theta) + \mu |\theta|_1 + \gamma |\theta|_2^2 \right], \ \mu, \lambda > 0.$$

For the LASSO, the 'optimal' theoretical choice of μ is proportional to $\sigma\sqrt{\log(p)/n}$ (see [8] for example). We see from Theorem 1 and 2 that the 'optimal' theoretical choice of λ for the exponential procedures is of the order n/σ^2. Thus we take for our experiment

$$\Lambda = (n/\sigma^2) \times \mathcal{G}, \quad \Lambda' = \sqrt{\sigma^2 \log(p)/n} \times \mathcal{G},$$

where $\mathcal{G}$ is defined as follows:

$$\mathcal{G} = 0.01 \times \{(1.5)^i, i = 1, ..., 15\} \cup \{0\}.$$

We compute for our exponential weights estimators the oracle quadratic risk $\inf_{\lambda \in \Lambda} |X(\hat{\theta}_n(\lambda, \pi) - \bar{\theta})|_2^2$ and $\inf_{\lambda \in \Lambda} |X(\tilde{\theta}_n(\lambda, \mathrm{m}) - \bar{\theta})|_2^2$, and similarly the oracle quadratic risks for the Lasso and the Elastic Net $\inf_{\mu \in \Lambda'} |X(\hat{\theta}_n^L(\mu) - \bar{\theta})|_2^2$ and $\inf_{\mu, \gamma \in \Lambda'} |X(\hat{\theta}_n^{EN}(\mu, \gamma) - \bar{\theta})|_2^2$.

5.2. **Numerical results.** We perform every experiment 20 times and give the results for the sparse situation in Table 1 and for the approximately sparse situation in Table 2. Convergence of the estimators can be checked on Figure 1.

We can see on these experiments that the exponential weights estimators outperforms the LASSO and the Elastic Net in the low noise case $\sigma = 1$. When σ grows, the performances of our estimators are still better, but the difference is less significative; moreover, $\hat{\theta}_n$ seems to become less stable (in particular Table 1, $p = 30$ and $\sigma^2 = 3$).

This simulation study clearly shows the advantage to use the exponential weights estimators, in particular $\tilde{\theta}_n$, especially in the situation of approximate sparsity. As we mentioned in the introduction, the main advantage of the LASSO and the Elastic

TABLE 1. Numerical results for the estimation of $\theta = (3, 1.5, 0, 0, 2, 0, 0, 0, 0, 0, ...)$. For each possible combination of σ and p, we report the median, mean and the standard deviation values of respectively $\inf_{\lambda\in\Lambda} |X(\hat{\theta}_n(\lambda, \pi) - \bar{\theta})|_2^2$, $\inf_{\lambda\in\Lambda} |X(\tilde{\theta}_n(\lambda, \mathrm{m}) - \bar{\theta})|_2^2$, $\inf_{\mu\in\Lambda'} |X(\hat{\theta}_n^L(\mu) - \bar{\theta})|_2^2$ and $\inf_{\mu,\gamma\in\Lambda'} |X(\hat{\theta}_n^{EN}(\mu, \gamma) - \bar{\theta})|_2^2$.

σ^2	p	what?	$\hat{\theta}_\mu^L$	$\hat{\theta}_{\mu,\gamma}^{EN}$	$\hat{\theta}_n$	$\tilde{\theta}_n$
1	8	median	0.302	0.302	0.176	**0.172**
		mean	0.291	0.291	0.215	**0.209**
		s.d.	0.211	0.211	0.190	0.185
3	8	median	0.437	0.437	0.533	**0.370**
		mean	0.535	**0.517**	0.612	0.527
		s.d.	0.398	0.388	0.420	0.395
1	30	median	0.355	0.355	0.157	**0.143**
		mean	0.360	0.358	0.217	**0.209**
		s.d.	0.189	0.188	0.151	0.150
3	30	median	1.459	1.459	1.511	**1.267**
		mean	1.431	1.408	1.809	**1.333**
		s.d.	0.702	0.690	1.143	0.607
1	100	median	0.399	0.399	0.244	**0.204**
		mean	0.471	0.471	0.248	**0.212**
		s.d.	0.222	0.222	0.162	0.130
3	100	median	1.378	**1.374**	1.674	1.409
		mean	1.396	1.395	1.800	**1.365**
		s.d.	0.687	0.688	0.653	0.562

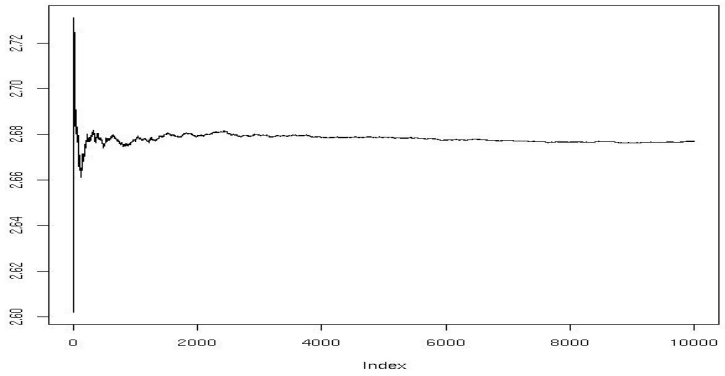

FIGURE 1. Convergence of the first coefficient of $\hat{\theta}_n$ in an experiment with $\sigma = 1$, $p = 8$. We represent the first coordinate of $\frac{1}{N-bo+1}\sum_{t=bo}^{N} \hat{\theta}_{I^{(t)}}$ as a function of $N = b_0, ..., T$.

Net is the computation time. When p becomes larger ($p > 1000$), the MCMC takes much longer to converge and the computation time becomes prohibitive. The strength of ℓ_1-penalized estimators is that they can be computed for large values of the dimension $p \simeq 10^7$ in a reasonable amount of time. On the other hand, these penalized methods are inferior to the exponential weights procedures in term of statistical properties when we consider the prediction problem.

5.3. **Some comments on computation time.** We can roughly analyze the computational complexity of our MCMC algorithm $\hat{\theta}_n$ as follows:

TABLE 2. Results for $\theta = 5 \times (e^{-1}, e^{-2}, ..., e^{-p})$. For each possible combination of σ and p, we report the median, mean and the standard deviation values of respectively $\inf_{\lambda\in\Lambda} |X(\hat{\theta}_n(\lambda, \pi) - \bar{\theta})|_2^2$, $\inf_{\lambda\in\Lambda} |X(\tilde{\theta}_n(\lambda, \mathrm{m}) - \bar{\theta})|_2^2$, $\inf_{\mu\in\Lambda'} |X(\hat{\theta}_n^L(\mu) - \bar{\theta})|_2^2$ and $\inf_{\mu,\gamma\in\Lambda'} |X(\hat{\theta}_n^{EN}(\mu, \gamma) - \bar{\theta})|_2^2$.

σ^2	p	what?	$\hat{\theta}_\mu^L$	$\hat{\theta}_{\mu,\gamma}^{EN}$	$\hat{\theta}_n$	$\tilde{\theta}_n$
1	8	median	0.138	0.138	**0.089**	0.121
		mean	0.178	0.175	**0.137**	**0.137**
		s.d.	0.145	0.145	0.121	0.116
3	8	median	0.397	0.397	0.437	**0.364**
		mean	0.434	0.414	0.430	**0.400**
		s.d.	0.178	0.286	0.271	0.282
1	30	median	0.262	0.262	**0.203**	0.205
		mean	0.277	0.276	0.247	**0.240**
		s.d.	0.147	0.147	0.149	0.149
3	30	median	0.593	0.593	0.519	**0.423**
		mean	0.630	0.619	0.665	**0.534**
		s.d.	0.409	0.420	0.684	0.383
1	100	median	0.276	0.271	**0.256**	0.261
		mean	0.375	0.375	0.353	**0.342**
		s.d.	0.256	0.256	0.200	0.199
3	100	median	1.045	1.045	0.687	**0.680**
		mean	1.023	1.023	0.809	**0.760**
		s.d.	0.364	0.363	0.476	0.464
1	1000	median	0.486	0.486	**0.390**	0.407
		mean	0.464	0.464	**0.373**	0.386
		s.d.	0.207	0.207	0.108	0.103
3	1000	median	1.549	1.547	**1.199**	1.288
		mean	1.483	1.481	1.268	**1.245**
		s.d.	0.460	0.458	0.702	0.692

(1) Fix T the number of MCMC steps.
(2) At each step $t \leq T$, we have to choose which new component we want to add (or remove) from the current model. There are at most p possible choices, and for each choice j we have to compute the correlation between the vectors $(Y_i - f_{\hat{\theta}_{J^{(t)}}})_{1\leq i\leq n}$ and $(\phi_j(X_i))_{1\leq i\leq n}$, this takes $\mathcal{O}(np)$ operations.
(3) Finally, at each step t we have to compute $\hat{\theta}_{J^{(t)}}$, this takes at most $\mathcal{O}(|J|^3)$ operations.

Finally, the number of operations is $\mathcal{O}\left(T(np + E_\lambda[|J|^3])\right)$ where $E_\lambda[|J|]$ is the expectation of $|J|$ under the aggregation distribution with temperature parameter λ

$$E_\lambda[|J|] = \frac{\sum_{J\in\mathcal{P}_n(\{1,...,p\})} \pi_J e^{-\lambda\left(r(\hat{\theta}_J) + \frac{2\sigma^2|J|}{n}\right)} |J|}{\sum_{J\in\mathcal{P}_n(\{1,...,p\})} \pi_J e^{-\lambda\left(r(\hat{\theta}_J) + \frac{2\sigma^2|J|}{n}\right)}}.$$

For properly tuned λ, we observe $\mathbb{E}[|J|] \simeq |J(\bar{\theta})|$. We understand why the sparsity (or approximate sparsity) of the parameter has an important influence on the computation time. Consider for example the case $p = 100$ and $n = 50$. If $|J(\bar{\theta})| = 10$ then $n \times p = 5000 > 1000 = |J(\bar{\theta})|^3$ whereas if $|J(\bar{\theta})| = 25$ then $n \times p = 5000 < 15625 = |J(\bar{\theta})|^3$.

All the simulations were performed with the R software [45]. The code are available on request by e-mail.

6. Proofs

6.1. **Proofs of Section 2.** This proof uses an argument from Leung and Barron [36].

Proof of Proposition 1. The mapping $Y \to \hat{f}_n(Y) \stackrel{\triangle}{=} (\hat{f}_n(X_1,Y),\dots,\hat{f}_n(X_n,Y))^T$ is clearly continuously differentiable by composition of elementary differentiable functions. For any subset $J \subset \{1,\dots,p\}$ define $A_J = (\phi_j(X_i))_{1\leq i\leq n, j\in J}$, $\Sigma_J = \frac{1}{n}A_J^T A_J$, $\Phi_J(\cdot) = (\phi_j(\cdot))_{j\in J}$ and

$$g_J = e^{-\lambda\left(\|Y-f_J\|_n^2 + \frac{2\sigma^2|J|}{n}\right)}$$

where

$$f_J(x,Y) = \frac{1}{n}Y^T A_J \Sigma_J^+ \Phi_J(x)^T,$$

and Σ_J^+ denotes the pseudo-inverse of Σ_J. Denote by ∂_i the derivative w.r.t. Y_i. Simple computations give

$$\partial_i f_J(x,Y) = \frac{1}{n}\Phi_J(X_i)\Sigma_J^+\Phi_J(x)^T,$$

$$(\partial_i f_J(X_1,Y),\dots,\partial_i f_J(X_n,Y))Y = f_J(X_i,Y),$$

and

$$\sum_{l=1}^n f_J(X_l,Y)\partial_i f_J(X_l,Y) \quad = \quad f_J(X_i,Y).$$

Thus we have

$$\begin{aligned}
\partial_i(g_J) &= -\lambda\partial_i\left(\|Y-f_J\|_n^2\right) g_J \\
&= -\frac{2\lambda}{n}\left((Y_i - f_J(X_i,Y)) - \sum_{l=1}^n \partial_i f_J(X_l,Y)(Y_l - f_J(X_l,Y))\right) g_J \\
&= -\frac{2\lambda}{n}(Y_i - f_J(X_i,Y))g_J,
\end{aligned}$$

Recall that

$$\hat{f}_n(\cdot,Y)) = \frac{\sum_{J\in\mathcal{P}_n(\{1,\dots,p\})}\pi_J g_J f_J(\cdot,Y)}{\sum_{J\in\mathcal{P}_n(\{1,\dots,p\})}\pi_J g_J}.$$

We have

$$\begin{aligned}
\partial_i\hat{f}_n(X_i,Y) &= \frac{\sum_{J\in\mathcal{P}_n(\{1,\dots,p\})}\pi_J\left(\partial_i(g_J)f_J(X_i,Y) + g_J\partial_i(f_J(X_i,Y))\right)}{\sum_{J\in\mathcal{P}_n(\{1,\dots,p\})}\pi_J g_J} \\
&\quad - \frac{\left(\sum_{J\in\mathcal{P}_n(\{1,\dots,p\})}\pi_J g_J f_J(X_i,Y)\right)\left(\sum_{J\in\mathcal{P}_n(\{1,\dots,p\})}\pi_J\partial_i(g_J)\right)}{(\sum_{J\in\mathcal{P}_n(\{1,\dots,p\})}\pi_J g_J)^2} \\
&= -\frac{2\lambda}{n}Y_i\hat{f}_n + \frac{2\lambda}{n}\frac{\sum_{J\in\mathcal{P}_n(\{1,\dots,p\})}f_J(X_i,Y)^2\pi_J g_J}{\sum_{J\in\mathcal{P}_n(\{1,\dots,p\})}\pi_J g_J} \\
&\quad + \frac{1}{n}\frac{\sum_{J\in\mathcal{P}_n(\{1,\dots,p\})}\Phi_J(X_i)\Sigma_J^+\Phi_J(X_i)^T\pi_J g_J}{\sum_{J\in\mathcal{P}_n(\{1,\dots,p\})}\pi_J g_J} \\
&+ \frac{2\lambda}{n}Y_i\hat{f}_n(X_i,Y) - \frac{2\lambda}{n}\hat{f}_n^2(X_i,Y)
\end{aligned}$$

$$= \frac{2\lambda}{n}\frac{\sum_{J\in\mathcal{P}_n(\{1,\dots,p\})}(f_J(X_i,Y)-\hat{f}_n(X_i,Y))^2\pi_J g_J}{\sum_{J\in\mathcal{P}_n(\{1,\dots,p\})}\pi_J g_J}$$

$$+\frac{1}{n}\frac{\sum_{J\in\mathcal{P}_n(\{1,\dots,p\})}\Phi_J(X_i)\Sigma_J^+\Phi_J(X_i)^T\pi_J g_J}{\sum_{J\in\mathcal{P}_n(\{1,\dots,p\})}\pi_J g_J}\geq 0. \tag{6.1}$$

Consider the following estimator of the risk

$$\hat{r}_n(Y)=\|\hat{f}_n(Y)-Y\|_n^2+\frac{2\sigma^2}{n}\sum_{i=1}^n\partial_i\hat{f}_n(X_i,Y)-\sigma^2. \tag{6.2}$$

Using an argument based on Stein's identity as in [35] we now prove that

$$\mathbb{E}[\hat{r}_n(Y)]=\mathbb{E}\left[\|\hat{f}_n(Y)-f\|_n^2\right].$$

We have

$$\begin{aligned}\mathbb{E}\left[\|\hat{f}_n(Y)-f\|_n^2\right] &= \mathbb{E}\left[\|\hat{f}_n(Y)-Y\|_n^2+\frac{2}{n}\sum_{i=1}^n W_i(\hat{f}_n(X_i,Y)-f(X_i))\right]-\sigma^2\\ &= \mathbb{E}\left[\|\hat{f}_n(Y)-Y\|_n^2+\frac{2}{n}\sum_{i=1}^n W_i\hat{f}_n(X_i,Y)\right]-\sigma^2. \qquad (6.3)\end{aligned}$$

For $\mathbf{z}=(z_1,\dots,z_n)^T\in\mathbb{R}^n$ write $F_{W,i}(\mathbf{z})=\prod_{j\neq i}F_{W,i}(z_j)$, where F_W denotes the c.d.f. of the random variable W_1. Since $\mathbb{E}(W_i)=0$ we have

$$\begin{aligned}\mathbb{E}\left[W_i\hat{f}_n(X_i,Y)\right] &= \mathbb{E}\left[W_i\int_0^{W_i}\partial_i\hat{f}_n(X_i,Y_1,\dots,Y_{i-1},f(X_i)+z,Y_{i+1},\dots,Y_n)dz\right]\\ &= \int_{\mathbb{R}^{n-1}}\left(\int_{\mathbb{R}}y\int_0^y\partial_i\hat{f}_n(X_i,f+\mathbf{z})dz_i dF_W(y)\right)dF_{W,i}(\mathbf{z}). \qquad (6.4)\end{aligned}$$

In view of (6.1) we can apply Fubini's Theorem to the right-hand-side of (6.4). We obtain under the assumption $W\sim\mathcal{N}(0,\sigma^2)$ that

$$\begin{aligned}\int_{\mathbb{R}^+}\int_0^y\partial_i\hat{f}_n(X_i,f+\mathbf{z})dz_i dF_W(y) &= \int_{\mathbb{R}^+}\int_{z_i}^\infty y dF_W(y)\partial_i\hat{f}_n(X_i,f+\mathbf{z})dz_i\\ &= \int_{\mathbb{R}^+}\sigma^2\partial_i\hat{f}_n(X_i,f+\mathbf{z})dF_W(z_i),\end{aligned}$$

A Similar equality holds for the integral over $\mathbb{R}^-$. Thus we obtain

$$\mathbb{E}\left[W_i\hat{f}_n(X_i,Y)\right]=\sigma^2\mathbb{E}\left[\partial_i\hat{f}_n(X_i,Y)\right].$$

Combining (6.2), (6.3) and the above display gives

$$\mathbb{E}\left[\hat{r}_n(Y)\right]=\mathbb{E}\left[\|\hat{f}_n(Y)-f\|_n^2\right].$$

Since $\hat{f}_n(\cdot,Y)$ is the expectation of $f_J(\cdot,Y)$ w.r.t. the probability distribution $\propto g\cdot\pi$, we have

$$\|\hat{f}_n(\cdot,Y)-Y\|_n^2=\frac{\sum_{J\in\mathcal{P}_n(\{1,\dots,p\})}\left(\|f_J(\cdot,Y)-Y\|_n^2-\|f_J(\cdot,Y)-\hat{f}_n(Y)\|_n^2\right)g_J\pi_J}{\sum_{J\in\mathcal{P}_n(\{1,\dots,p\})}g_J\pi_J}.$$

For the sake of simplicity set $f_J=f_J(\cdot,Y)$ and $\hat{f}_n=\hat{f}_n(\cdot,Y)$. Combining (6.2), the above display and $\lambda\leqslant\frac{n}{4\sigma^2}$ yields

$$\hat{r}_n(Y) = \frac{\sum_{J\in\mathcal{P}_n(\{1,\dots,p\})}\left(\|f_J-Y\|_n^2+\sum_{i=1}^n\left(\frac{4\lambda\sigma^2}{n}-1\right)\|f_J-\hat{f}_n\|_n^2\right)g_J\pi_J}{\sum_{J\in\mathcal{P}_n(\{1,\dots,p\})}\pi_J g_J}$$

$$+\frac{2\sigma^2}{n^2}\sum_{i=1}^{n}\frac{\sum_{J\in\mathcal{P}_n(\{1,\dots,p\})}\Phi_J(X_i)\Sigma_J^+\Phi_J(X_i)^T\pi_J g_J}{\sum_{J\in\mathcal{P}_n(\{1,\dots,p\})}\pi_J g_J}-\sigma^2$$

$$\leqslant \sum_{J\in\mathcal{P}_n(\{1,\dots,p\})}\left(\|f_J-Y\|_n^2+\frac{2\sigma^2}{n}|J|\right)g_J\pi_J-\sigma^2.$$

By definition of g_J we have

$$\|f_J-Y\|_n^2+\frac{2\sigma^2|J|}{n}=-\frac{1}{\lambda}\log\left(\frac{g_J}{\sum_{J\in\mathcal{P}_n(\{1,\dots,p\})}g_J\pi_J}\right)-\frac{1}{\lambda}\log\left(\sum_{J\in\mathcal{P}_n(\{1,\dots,p\})}g_J\pi_J\right).$$

Integrating the above inequality w.r.t. the probability distribution $\frac{1}{C}g\cdot\pi$ (where $C=\sum_{J\in\mathcal{P}_n(\{1,\dots,p\})}g_J\pi_J$ is the normalization factor) and using the fact that

$$\sum_{J\in\mathcal{P}_n(\{1,\dots,p\})}\frac{1}{C}g_J\pi_J\log\left(\frac{1}{C}g_J\right)=K\left(\frac{g\cdot\pi}{C},\pi\right)\geqslant 0$$

as well as a convex duality argument (cf., e.g., [23], p. 264) we get

$$\hat{r}_n(Y)\leqslant\sum_{J\in\mathcal{P}_n(\{1,\dots,p\})}\left(\|Y-f_J\|_n^2+\frac{2\sigma^2}{n}|J|\right)\pi'_J+\frac{1}{\lambda}K(\pi',\pi)-\sigma^2,$$

for all probability measure π' on $\mathcal{P}(\{1,\dots,p\})$. Taking the expectation in the last inequality we get for any π'

$$\begin{aligned}
\mathbb{E}\left[\|\hat{f}_n-f\|_n^2\right] &= \mathbb{E}[\hat{r}_n(Y)]\\
&\leqslant \sum_{J\in\mathcal{P}_n(\{1,\dots,p\})}\left(\mathbb{E}[\|f_J-Y\|_n^2]+\frac{2\sigma^2}{n}|J|\right)\pi'_J+\frac{1}{\lambda}K(\pi',\pi)-\sigma^2\\
&\leqslant \sum_{J\in\mathcal{P}_n(\{1,\dots,p\})}\left(\mathbb{E}[\|f_J-f\|_n^2]+\frac{2}{n}\sum_{i=1}^{n}\mathbb{E}[W_i f_J(X_i,Y)]+\frac{2\sigma^2}{n}|J|\right)\pi'_J\\
&\qquad +\frac{1}{\lambda}K(\pi',\pi)\\
&\leqslant \sum_{J\in\mathcal{P}_n(\{1,\dots,p\})}\left(\mathbb{E}[\|f_J-f\|_n^2]+\frac{4\sigma^2}{n}|J|\right)\pi'_J+\frac{1}{\lambda}K(\pi',\pi),
\end{aligned}$$

where we have used Stein's argument $\mathbb{E}[W_i f_J(X_i,Y)]=\sigma^2\mathbb{E}\left[\partial_i f_J(X_i,Y)\right]$ and the fact that $\sum_{i=1}^{n}\partial_i f_J(X_i,Y)=1$ in the last line. Finally taking π' in the set of Dirac distributions on the subset J of $\{1,\dots,p\}$ yields the theorem. □

Proof of Theorem 1. First note that any minimizer $\theta\in\mathbb{R}^p$ of the right-hand-side in (2.6) is such that $|J(\theta)|\leqslant\mathrm{rank}(A)\leqslant n$ where we recall that $A=(\phi_j(X_i))_{1\leqslant i\leqslant n,1\leqslant j\leqslant p}$. Indeed, for any $\theta\in\mathbb{R}^p$ such that $|J(\theta)|>\mathrm{rank}(A)$ we can construct a vector $\theta'\in\mathbb{R}^p$ such that $f_\theta=f_{\theta'}$ and $|J(\theta')|\leqslant\mathrm{rank}(A)$ and the mapping $x\to x\log\left(\frac{ep\alpha}{x}\right)$ is non-decreasing on $(0,p]$.

Next for any $J\in\mathcal{P}_n(\{1,\dots,p\})$ we have

$$\mathbb{E}[\|f_J-f\|_n^2]=\min_{\theta\in\Theta(J)}\left\{\|f_\theta-f\|_n^2\right\}+\frac{\sigma^2|J|}{n}=\min_{\theta\in\Theta(J)}\left\{\|f_\theta-f\|_n^2+\frac{\sigma^2|J(\theta)|}{n}\right\}.$$

Thus

$$\begin{aligned}
&\min_{J\in\mathcal{P}_n(\{1,\dots,p\})}\left\{\mathbb{E}[\|f_J-f\|_n^2]+\frac{1}{\lambda}\log\left(\frac{1}{\pi_J}\right)+\frac{\sigma^2 J}{n}\right\}\\
&\qquad=\min_{J\in\mathcal{P}_n(\{1,\dots,p\})}\min_{\theta\in\Theta(J)}\left\{\|f_\theta-f\|_n^2+\frac{1}{\lambda}\log\left(\frac{1}{\pi_{J(\theta)}}\right)+\frac{\sigma^2|J(\theta)|}{n}\right\}
\end{aligned}$$

$$= \min_{\theta\in\mathbb{R}^p} \left\{ \|f_\theta - f\|_n^2 + \frac{1}{\lambda}\log\left(\frac{1}{\pi_{J(\theta)}}\right) + \frac{\sigma^2|J(\theta)|}{n} \right\}.$$

Combining the above display with Proposition 1 and our definition of the prior π gives the result. □

6.2. **Proof of Theorem 2.** We state below a version of Bernstein's inequality useful in the proof of Theorem 2. See Proposition 2.9 page 24 in [41], more precisely Inequality (2.21).

Lemma 1. *Let T_1, ..., T_n be independent real valued random variables. Let us assume that there is two constants v and w such that*

$$\sum_{i=1}^n \mathbb{E}[T_i^2] \le v$$

and for all integers $k \ge 3$,

$$\sum_{i=1}^n \mathbb{E}\left[(T_i)_+^k\right] \le v\frac{k!w^{k-2}}{2}.$$

Then, for any $\zeta \in (0, 1/w)$,

$$\mathbb{E}\exp\left[\zeta\sum_{i=1}^n [T_i - \mathbb{E}(T_i)]\right] \le \exp\left(\frac{v\zeta^2}{2(1-w\zeta)}\right).$$

Proof of Theorem 2. For any $\theta \in \Theta_{K+c}$ define the random variables

$$T_i = T_i(\theta) = -\left(Y_i - f_\theta(X_i)\right)^2 + \left(Y_i - f_{\bar\theta}(X_i)\right)^2.$$

Note that these variables are independent. We have

$$\begin{aligned}
\sum_{i=1}^n \mathbb{E}[T_i^2] &= \sum_{i=1}^n \mathbb{E}\left[\left[2Y_i - f_{\bar\theta}(X_i) - f_\theta(X_i)\right]^2 \left[f_{\bar\theta}(X_i) - f_\theta(X_i)\right]^2\right] \\
&= \sum_{i=1}^n \mathbb{E}\left[\left[2W_i + 2f(X_i) - f_{\bar\theta}(X_i) - f_\theta(X_i)\right]^2 \left[f_{\bar\theta}(X_i) - f_\theta(X_i)\right]^2\right] \\
&\le \sum_{i=1}^n \mathbb{E}\left[\left[8W_i^2 + 2(2\|f\|_\infty + L(2K+c))^2\right] \left[f_{\bar\theta}(X_i) - f_\theta(X_i)\right]^2\right] \\
&= \sum_{i=1}^n \mathbb{E}\left[8W_i^2 + 2(2\|f\|_\infty + L(2K+c))^2\right] \mathbb{E}\left[\left[f_{\bar\theta}(X_i) - f_\theta(X_i)\right]^2\right] \\
&\le n\left[8\sigma^2 + 2(2\|f\|_\infty + L(2K+c))^2\right]\left[R(\theta) - R(\bar\theta)\right] =: v(\theta,\bar\theta) = v,
\end{aligned}$$

where we have used in the last line that $\|f_\theta - f_{\bar\theta}\|^2 = R(\theta) - R(\bar\theta)$. Next we have, for any integer $k \ge 3$, that

$$\begin{aligned}
\sum_{i=1}^n \mathbb{E}\left[(T_i)_+^k\right] &\le \sum_{i=1}^n \mathbb{E}\left[|2Y_i - f_{\bar\theta}(X_i) - f_\theta(X_i)|^k \, |f_{\bar\theta}(X_i) - f_\theta(X_i)|^k\right] \\
&\le \sum_{i=1}^n \mathbb{E}\left[2^{2k-1}\left[|W_i|^k + (\|f\|_\infty + L(K+c/2)^k\right] |f_{\bar\theta}(X_i) - f_\theta(X_i)|^k\right] \\
&\le \sum_{i=1}^n \mathbb{E}\left[2^{2k-1}\left[|W_i|^k + (\|f\|_\infty + L(K+c/2)^k\right] [L(2K+c)]^{k-2} \left[f_{\bar\theta}(X_i) - f_\theta(X_i)\right]^2\right] \\
&\le 2^{2k-1}\left[\sigma^2 k!\xi^{k-2} + (\|f\|_\infty + L(K+c/2))^k\right][L(2K+c)]^{k-2} \sum_{i=1}^n \mathbb{E}\left[\left[f_{\bar\theta}(X_i) - f_\theta(X_i)\right]^2\right]
\end{aligned}$$

$$\leq \frac{(\sigma^2 k!\xi^{k-2} + (\|f\|_\infty + L(K+c/2))^k)(4L(2K+c))^{k-2}}{4(\sigma^2 + (\|f\|_\infty + L(K+c/2))^2)} v$$
$$\leq \frac{1}{4}\left(k!\xi^{k-2} + [\|f\|_\infty + L(K+c/2)]^{k-2}\right)[4L(2K+c)]^{k-2} v$$
$$\leq \frac{2}{4} k! \left(\xi + [\|f\|_\infty + L(K+c/2)]\right)^{k-2} [4L(2K+c)]^{k-2v} \leq v \frac{k! w^{k-2}}{2},$$

with $w := 8(\xi + [\|f\|_\infty + L(K+c/2)])L(K+c/2)$.

Next, for any $\lambda \in (0, n/w)$ and $\theta \in \Theta_{K+c}$, applying Lemma 1 with $\zeta = \lambda/n$ gives

$$\mathbb{E}\exp\left[\lambda\Big(R(\theta) - R(\bar\theta) - r(\theta) + r(\bar\theta)\Big)\right] \leq \exp\left[\frac{v\lambda^2}{2n^2(1-\frac{w\lambda}{n})}\right].$$

Set $C = 8\left(\sigma^2 + [\|f\|_\infty + L(K+c/2)]^2\right)$. For any $\varepsilon > 0$ the last display yields

$$\mathbb{E}\exp\left[\left(\lambda - \frac{\lambda^2 C}{2n(1-\frac{w\lambda}{n})}\right)\Big(R(\theta) - R(\bar\theta)\Big) + \lambda\Big(-r(\theta) + r(\bar\theta)\Big) - \log\frac{2}{\varepsilon}\right] \leq \frac{\varepsilon}{2}.$$

Integrating w.r.t. the probability distribution $\mathrm{m}(\cdot)$ we get

$$\int \mathbb{E}\exp\left[\left(\lambda - \frac{\lambda^2 C}{2n(1-\frac{w\lambda}{n})}\right)\Big(R(\theta) - R(\bar\theta)\Big) + \lambda\Big(-r(\theta) + r(\bar\theta)\Big) - \log\frac{2}{\varepsilon}\right]\mathrm{m}(d\theta) \leq \frac{\varepsilon}{2}.$$

Next, Fubini's theorem gives

$$\mathbb{E}\int \exp\left[\left(\lambda - \frac{\lambda^2 C}{2n(1-\frac{w\lambda}{n})}\right)\Big(R(\theta) - R(\bar\theta)\Big) + \lambda\Big(-r(\theta) + r(\bar\theta)\Big) - \log\frac{2}{\varepsilon}\right]\mathrm{m}(d\theta) \leq \frac{\varepsilon}{2}.$$

$$\mathbb{E}\int \exp\left[\left(\lambda - \frac{\lambda^2 C}{2n(1-\frac{w\lambda}{n})}\right)\Big(R(\theta) - R(\bar\theta)\Big) + \lambda\Big(-r(\theta) + r(\bar\theta)\Big) - \log\left[\frac{d\tilde\rho_\lambda}{d\mathrm{m}}(\theta)\right] - \log\frac{2}{\varepsilon}\right]\tilde\rho_\lambda(d\theta) \leq \frac{\varepsilon}{2}.$$

Jensen's inequality yields

$$\mathbb{E}\exp\left[\left(\lambda - \frac{\lambda^2 C}{2n(1-\frac{w\lambda}{n})}\right)\left(\int R d\tilde\rho_\lambda - R(\bar\theta)\right) + \lambda\left(-\int r d\tilde\rho_\lambda + r(\bar\theta)\right) - \mathcal{K}(\tilde\rho_\lambda, \mathrm{m}) - \log\frac{2}{\varepsilon}\right] \leq \frac{\varepsilon}{2}.$$

Now, using the basic inequality $\exp(x) \geq \mathbb{1}_{\mathbb{R}_+}(x)$ we get

$$\mathbb{P}\left\{\left(\lambda - \frac{\lambda^2 C}{2n(1-\frac{w\lambda}{n})}\right)\left(\int R d\tilde\rho_\lambda - R(\bar\theta)\right) + \lambda\left(-\int r d\tilde\rho_\lambda + r(\bar\theta)\right) - \mathcal{K}(\tilde\rho_\lambda, \mathrm{m}) - \log\frac{2}{\varepsilon} \geq 0\right\} \leq \frac{\varepsilon}{2}.$$

Using Jensen's inequality again gives

$$\int Rd\tilde{\rho}_\lambda \geq R\left(\int \theta\tilde{\rho}_\lambda(d\theta)\right) = R(\tilde{\theta}_\lambda).$$

Combining the last two displays we obtain

$$\mathbb{P}\left\{R(\tilde{\theta}_\lambda) - R(\bar{\theta}) \leq \frac{\int rd\tilde{\rho}_\lambda - r(\bar{\theta}) + \frac{1}{\lambda}\left[\mathcal{K}(\tilde{\rho}_\lambda, \mathrm{m}) + \log\frac{2}{\varepsilon}\right]}{1 - \frac{\lambda C}{2(n-w\lambda)}}\right\} \geq 1 - \frac{\varepsilon}{2}.$$

Now, using Lemma 1.1.3 in Catoni [16] we obtain that

(6.5)
$$\mathbb{P}\left\{R(\tilde{\theta}_\lambda) - R(\bar{\theta}) \leq \inf_{\rho\in\mathcal{M}_+^1(\Theta_{K+c})} \frac{\int rd\rho - r(\bar{\theta}) + \frac{1}{\lambda}\left[\mathcal{K}(\rho, \mathrm{m}) + \log\frac{2}{\varepsilon}\right]}{1 - \frac{\lambda C}{2(n-w\lambda)}}\right\} \geq 1 - \frac{\varepsilon}{2}.$$

We now want to bound from above $r(\theta) - r(\bar{\theta})$ by $R(\theta) - R(\bar{\theta})$. Applying Lemma 1 to $\tilde{T}_i(\theta) = -T_i(\theta)$ and similar computations as above yield successively

$$\mathbb{E}\exp\left[\lambda\Big(R(\bar{\theta}) - R(\theta) + r(\theta) - r(\bar{\theta})\Big)\right] \leq \exp\left[\frac{v\lambda^2}{2n^2(1 - \frac{w\lambda}{n})}\right],$$

and so for any (data-dependent) ρ,

$$\mathbb{E}\exp\Bigg[\left(\lambda + \frac{\lambda^2 C}{2(n-w\lambda)}\right)\left(-\int Rd\rho + R(\bar{\theta})\right) + \lambda\left(\int rd\rho - r(\bar{\theta})\right) - \mathcal{K}(\rho, \mathrm{m}) - \log\frac{2}{\varepsilon}\Bigg] \leq \frac{\varepsilon}{2},$$

and

(6.6)
$$\mathbb{P}\Bigg\{\int rd\rho - r(\bar{\theta}) \leq \left(1 + \frac{\lambda C}{2(n-w\lambda)}\right)\left[\int Rd\rho - R(\bar{\theta})\right] + \frac{1}{\lambda}\left[\mathcal{K}(\rho, \mathrm{m}) + \log\frac{2}{\varepsilon}\right]\Bigg\} \geq 1 - \frac{\varepsilon}{2}.$$

Combining (6.6) and (6.5) with a union bound argument gives

$$\mathbb{P}\Bigg\{R(\tilde{\theta}_\lambda) - R(\bar{\theta}) \leq \inf_{\rho\in\mathcal{M}_+^1(\Theta_{K+c})} \frac{\left(1 + \frac{\lambda C}{2(n-w\lambda)}\right)\left[\int Rd\rho - R(\bar{\theta})\right] + \frac{2}{\lambda}\left[\mathcal{K}(\rho, \mathrm{m}) + \log\frac{2}{\varepsilon}\right]}{1 - \frac{\lambda C}{2(n-w\lambda)}}\Bigg\} \geq 1 - \varepsilon,$$

where $\mathcal{M}_+^1(\Theta_{K+c})$ is the set of all probability measures over Θ_{K+c}.

Now for any $\delta \in (0, c]$ taking ρ as the uniform probability measure on the set $\{t \in \Theta(J(\bar{\theta})) : |t - \bar{\theta}|_1 \leq \delta\} \subset \Theta_{K+c}(J(\bar{\theta}))$ gives

$$\mathbb{P}\Bigg\{R(\tilde{\theta}_\lambda) \leq R(\bar{\theta}) + \frac{1}{1 - \frac{\lambda C}{2(n-w\lambda)}}\Bigg[\left(1 + \frac{\lambda C}{2(n-w\lambda)}\right)L^2\delta^2 + \frac{2}{\lambda}\left(|J(\bar{\theta})|\log\frac{K+c}{\delta} + |J(\bar{\theta})|\log\frac{1}{\alpha} + \log\left(\frac{1}{1-\alpha}\right) + \log\binom{p}{|J(\bar{\theta})|} + \log\frac{2}{\varepsilon}\right)\Bigg]\Bigg\} \geq 1 - \varepsilon.$$

Taking $\delta = c = n^{-1}$ and the inequality $\log\left(\binom{p}{|J(\bar{\theta})|}\right) \leq |J(\bar{\theta})| \log \frac{pe}{|J(\bar{\theta})|}$ gives

$$\mathbb{P}\Bigg\{ R(\tilde{\theta}_\lambda) \leq R(\bar{\theta}) + \frac{1}{1-\frac{\lambda C}{2(n-w\lambda)}} \Bigg[\left(1 + \frac{\lambda C}{2(n-w\lambda)}\right) \frac{L^2}{n^2} + \frac{2}{\lambda}\left(|J(\bar{\theta})| \log(K+c) + |J(\bar{\theta})| \log\left(\frac{epn}{\alpha |J(\bar{\theta})|}\right) + \log\left(\frac{2}{\varepsilon(1-\alpha)}\right)\right)\Bigg]\Bigg\} \geq 1-\varepsilon \tag{6.7}$$

Taking now $\lambda = n/(2\mathcal{C}_1)$ (where we recall that $\mathcal{C}_1 = C \vee w$) in (6.7) gives

$$\mathbb{P}\Bigg\{ R(\tilde{\theta}_\lambda) \leq R(\bar{\theta}) + \frac{3L^2}{n^2} + \frac{8\mathcal{C}_1}{n}\Bigg[|J(\bar{\theta})| \log(K+c) + \left(|J(\bar{\theta})| \log\left(\frac{enp}{\alpha |J(\bar{\theta})|}\right) + \log\left(\frac{2}{\varepsilon(1-\alpha)}\right)\right)\Bigg]\Bigg\} \geq 1-\varepsilon,$$

where we have used that $1 - \frac{\lambda C}{2(n-w\lambda)} \geq 1/2$ and $1 + \frac{\lambda C}{2(n-w\lambda)} \leq 3/2$. □

References

[1] H. Akaike. Information theory and an extension of the maximum likelihood principle. In B. N. Petrov and F. Csaki, editors, *2nd International Symposium on Information Theory*, pages 267–281. Budapest: Akademia Kiado, 1973.

[2] P. Alquier. *Transductive and Inductive Adaptative Inference for Regression and Density Estimation*. PhD thesis, University Paris 6, 2006.

[3] P. Alquier. Pac-bayesian bounds for randomized empirical risk minimizers. *Mathematical Methods of Statistics*, 17(4):279–304, 2008.

[4] J.-Y. Audibert. Aggregated estimators and empirical complexity for least square regression. *Annales de l'Institut Henri Poincaré: Probability and Statistics*, 40(6):685–736, 2004.

[5] J.-Y. Audibert. *PAC-Bayesian Statistical Learning Theory*. PhD thesis, University Paris 6, 2004.

[6] F. Bach. Model-consistent sparse estimation through the bootstrap. 2009.

[7] A. Barron, A. Cohen, W. Dahmen, and R. DeVore. Adaptative approximation and learning by greedy algorithms. *The annals of statistics*, 36(1):64–94, 2008.

[8] P. J. Bickel, Y. Ritov, and A. Tsybakov. Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009.

[9] M. Bogdan, A. Chakrabarti, and J. K. Ghosh. Optimal rules for multiple testing and sparse multiple regression. Technical report I-18/08/P-003, Wroclaw University of Technology, 2008.

[10] F. Bunea, A. Tsybakov, and M. Wegkamp. Sparsity oracle inequalities for the lasso. *Electronic Journal of Statistics*, 1:169–194, 2007.

[11] F. Bunea, A.B. Tsybakov, and M.H. Wegkamp. Aggregation for Gaussian regression. *Annals of Statistics*, 35:1674–1697, 2007.

[12] T. Cai, G. Xu, and J. Zhang. On recovery of sparse signals via ℓ_1 minimization. *IEEE Transactions on Information Theory*, 55:3388–3397, 2009.

[13] E. Candes and T. Tao. The dantzig selector: statistical estimation when p is much larger than n. *The Annals of Statistics*, 35, 2007.

[14] O. Catoni. A pac-bayesian approach to adaptative classification. *Preprint Laboratoire de Probabilités et Modèles Aléatoires*, 2003.

[15] O. Catoni. *Statistical Learning Theory and Stochastic Optimization, Lecture Notes in Mathematics (Saint-Flour Summer School on Probability Theory 2001, ed. J. Picard)*. Springer, 2004.

[16] O. Catoni. *PAC-Bayesian Supervised Classification (The Thermodynamics of Statistical Learning)*, volume 56 of *Lecture Notes-Monograph Series*. IMS, 2007.

[17] S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM Review*, 43(1):129–159, 2001.

[18] W. Cui and I. E. George. Empirical bayes vs. fully bayes variable selection. *Journal of Statistical Planning and Inference*, 138(4):888–900, 2008.

[19] A. Dalalyan and A. Tsybakov. Aggregation by exponential weighting, sharp oracle inequalities and sparsity. *Machine Learning*, 72(1-2):39–61, 2008.

[20] A.S. Dalalyan and A.B. Tsybakov. Pac-bayesian bounds for the expected error of aggregation by exponential weights. Technical report, Université Paris 6, CREST and CERTIS, Ecole des Ponts ParisTech, 2009. personal communication.
[21] A.S. Dalalyan and A.B. Tsybakov. Mirror averaging with sparsity priors. arXiv:1003:1189, 2010.
[22] A.S. Dalalyan and A.B. Tsybakov. Sparse regression learning by aggregation and langevin monte-carlo. arXiv:09.1223v3, 2010.
[23] A. Dembo and O. Zeitouni. *Large Deviation Techniques and Applications*. Springer, 1998.
[24] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *Ann. Statist.*, 32(2):407–499, 2004.
[25] L. Frank and J. Friedman. A statistical view on some chemometrics regression tools. *Technometrics*, 16:499–511, 1993.
[26] I. E. George. The variable selection problem. *Journal of the American Statistician Association*, 95(452):1304–1308, 2000.
[27] I. E. George and R. E. McCulloch. Approaches for bayesian model selection. *Statistica Sinica*, 7:339–373, 1997.
[28] S. Ghosh. Adaptive elastic net: an improvement of elastic net to achieve oracle properties. Preprint, 2007.
[29] P. J. Green. Reversible jump markov chain monte carlo computation and bayesian model determination. *Biometrika*, 82(4):711–732, 1995.
[30] P. J. Green and S. Richardson. On bayesian analysis of mixtures with an unknown number of components. *Journal of the Royal Statistical Society: series B (statistical methodology)*, 59(4):731–792, 1997.
[31] C. Huang, G. L. H. Cheang, and A. Barron. Risk of penalized least squares, greedy selection and l1 penalization for flexible function libraries. Preprint, 2008.
[32] W. Jiang. Bayesian variable selection for high dimensionnal generalized linear models: Convergence rate of the fitted density. *The Annals of Statistics*, 35(4):1487–1511, 2007.
[33] Anatoli Juditsky, Philippe Rigollet, and Alexandre Tsybakov. Learning by mirror averaging. *Ann. Stat.*, 36(5):2183–2206, 2008.
[34] V. Koltchinskii. Sparsity in empirical risk minimization. Annales de l'Institut Henri Poincaré, Probability and Statistics (to appear).
[35] E. L. Lehmann and G. Casella. *Theory of Point Estimation*. Springer, 1998.
[36] G. Leung and A.R. Barron. Information theory and mixing least-squares regressions. *IEEE Trans. Inform. Theory*, 52(8):3396–3410, 2006.
[37] F. Liang, R. Paulo, G. Molina, M. Clyde, and J. O. Berger. Mixture of g-priors for bayesian variable selection. *Journal of the American Statistical Association*, 103:410–423, 2008.
[38] K. Lounici. Generalized mirror averaging and d-convex aggregation. *Mathematical Methods of Statistics*.
[39] C. L. Mallows. Some comments on c_p. *Technometrics*, 15:661–676, 1973.
[40] J.-M. Marin and C. Robert. *Bayesian Core: A practical approach to computational Bayesian analysis*. Springer, 2007.
[41] P. Massart. *Concentration Inequalities and Model Selection (Saint-Flour Summer School on Probability Theory 2003, ed. J. Picard)*. Springer, 2007.
[42] D. A. McAllester. Some pac-bayesian theorems. In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory (Madison, WI, 1998)*, pages 230–234. ACM, 1998.
[43] N. Meinshausen and P. Bühlmann. Stability selection. To appear in Journal of the Royal Statistical Society B, 2010.
[44] D. J. Nott and D. Leonte. Sampling schemes for bayesian variable selection in geleralized linear models. *Journal of Computationnal and Graphical Statistics*, 13:362–382, 2004.
[45] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2008. ISBN 3-900051-07-0.
[46] P. Rigollet and A.B. Tsybakov. Exponential screening and optimal rates of sparse estimation. manuscript. manuscript, 2010.
[47] G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6:461–464, 1978.
[48] J. G. Scott and J. O. Berger. Bayes and empirical-bayes multiplicity adjustement in the variable-selection problem. The Annals of Statistics, to appear, 2010.
[49] J. Shawe-Taylor and R. Williamson. A pac analysis of a bayes estimator. In *Proceedings of the Tenth Annual Conference on Computational Learning Theory, COLT'97*, pages 2–9. ACM, 1997.
[50] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society B*, 58(1):267–288, 1996.

[51] A.B. Tsybakov. Optimal rates of aggregation. In *Computationnal Learning theory and Kernel Machines (COLT)*, volume 2777 of *Lecture Notes in Artificial Intelligence*, pages 303–313. Springer, heidelberg, 2003.
[52] S. A. van de Geer and P. Bühlmann. On the conditions used to prove oracle results for the lasso. *Electronic Journal of Statistics*, 3:1360–1392, 2009.
[53] M. West. Bayesian factors in the "large p, small n" paradigm. *Bayesian statistics*, 7:723–732, 2003.
[54] Y. Yang. Aggregating regression procedures to improve performance. *Bernoulli*, 10(1):25–47, 2004.
[55] T. Zhang. From epsilon-entropy to kl-entropy: analysis of minimum information complexity density estimation. *The Annals of Statistics*, 34:2180–2210, 2006.
[56] H. Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101:1418–1429, 2006.
[57] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society B*, 67(2):301–320, 2005.

CREST and LPMA -Université Paris 7, School of Mathematics - Georgia Tech
E-mail address: alquier@ensae.fr, klounici@math.gatech.edu